A Large-scale Study of the Effect of Training Set Characteristics over Learning-to-Rank Algorithms

Evangelos Kanoulas
Stefan Savev, Pavel Metrikov, Virgil Pavlu and Javed Aslam

Consider the problem of the low-cost construction of effective training datasets with the respect to learning-to-rank algorithms.

What effect has the distribution of labels across the different grades of relevance in the training set on the performance of the learning to rank algorithms?

Data

“WEB30k” dataset by Microsoft Research:
• ~30,000 web queries
• 136 features for each query-url pair
• 5 grades of relevance

Methodology

1. Construct thousands of judgment sets (points on Figure 1), each one having some pre-defined distribution of labels:
   • combine the grades “3” and “4”
   • select only queries that have at least k labeled urls in each of the relevance grades

2. Run learning-to-rank algorithms over the judgment sets.
3. Measure the performance in terms of nDCG@10.

Results

Informative summaries found:
1. The normalized cumulative gain of the judgment set
2. The variance over the judgment set

Conclusions

1. Distributions with a balance between the number of documents in the extreme grades are to be favored
2. The middle relevance grades play less important role than the extreme ones

References